DOCUMENT RESUME

ED 338 707 TM 017 572

AUTHOR Sireci, Stephen G.
TITLE "Sample-Independent" Item Parameters? An Investigation of the Stability of IRT Item Parameters Estimated from Small Data Sets.
PUB DATE 24 Oct 91
NOTE 29p.; Paper presented at the Annual Meeting of the Northeastern Educational Research Association (Ellenville, NY, October 1991).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Certified Public Accountants; Computer Simulation; Data Analysis; *Estimation (Mathematics); Evaluation Methods; *Item Response Theory; Licensing Examinations (Professions); Research Methodology; *Sample Size; *Test Reliability
IDENTIFIERS Biserial Correlation; Classical Test Theory; *Data Sets; Invariance Principle; Item Parameters; One Parameter Model; P Values; Three Parameter Model; Two Parameter Model; *Unidimensionality (Tests)

ABSTRACT

Whether item response theory (IRT) is useful to the small-scale testing practitioner is examined. The stability of IRT item parameters is evaluated with respect to the classical item parameters (i.e., p-values, biserials) obtained from the same data set. Previous research investigating the effect of sample size on IRT parameter estimation has usually been performed on simulated data. The present study follows a few others in using a real-life small-sample testing application. The procedure involved obtaining a common set of items administered to three small-sample groups of examinees over a 3-year period (sample sizes of 173 in 1988, 149 in 1989, and 106 in 1990, respectively) and estimating the item parameters with both restricted and unrestricted IRT models. The test was a national certification examination for certified public accountants wanting certification in personal financial counseling. Classical reliability (p-value) and biserial correlation analyses were performed prior to traditional one-, two-, and three-parameter IRT analyses. Results suggest that stable item difficulty parameters can be obtained for small sample sizes using the one-parameter or modified two-parameter model when the data fit the IRT model (i.e., when they are unidimensional). The IRT and classical analyses performed could not successfully provide stable item discrimination parameters. However, the conditions under which IRT is useful to the small-sample test practitioner are discussed. Eleven tables present study data. A 24-item list of references is included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED338707

"Sample-Independent" Item Parameters?

An Investigation of the Stability of IRT Item Parameters Estimated from Small Data Sets

Stephen G. Sireci[1]

American Institute of Certified Public Accountants

Examinations Division

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

STEPHEN G. SIRECI

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

A paper presented at the Annual Conference of the Northeastern Eduational Research Association, Ellenville, New York, October 24, 1991

---

[1]The author thanks Lynn Shelley for her helpful suggestions concerning the research design, David Thissen for use of the Multilog program and for assistance in running the restricted IRT models, and Howard Wainer for sharing his insights regarding the assessment of group differences using IRT. The author maintains the sole responsibility for the content of this paper and any errors therein. The views expressed in this paper do not reflect an official position of the American Institute of Certified Public Accountants.

TM017572

BEST COPY AVAILABLE

Statement of the Problem

The benefits of using Item Response Theory (IRT) to develop, score and evaluate tests have been widely espoused. Proponents of IRT claim that it has several advantages over classical test theory (e.g., Hambleton, 1989). Perhaps the greatest advantage of IRT is the claim that the person and item parameters obtained using IRT are *sample independent*. This claim states that IRT parameters are independent of the particular sample of items and/or examinees chosen. Thus, item parameters obtained from one group of examinees should remain stable across other groups of examinees, and person parameter estimates should remain stable (invariant) across other groups of test items[1].

A problem with IRT is that a large number of examinees are often required to obtain stable item parameters (Hambleton, 1989; Hulin, Lissak, & Drasgow, 1982; Thissen & Wainer, 1982); however, in many testing situations, tests are administered to relatively small numbers of examinees. Thus, the sample size requirements of IRT often preclude its use in small-sample testing applications. The purpose of this paper is to discover if IRT can be useful to the small-scale testing practitioner. Because these practitioners often use classical item parameters (i.e., p-values, biserials), the stability of the IRT item parameters will be evaluated with respect to the classical parameters obtained from the same data set.

---

[1]This claim assumes that the different groups of examinees are not widely different in terms of the attribute being measured and that the groups of items are selected from the same item pool.

Previous Studies of Item Parameter Stability

Previous research has indicated that the minimum number of examinees recommended for accurate item parameter estimation varies widely according to the specific testing situation and the particular IRT model chosen (Barnes & Wise, 1991; Wainer & Mislevy, 1990). In general, longer tests and more complex (general) IRT models require larger sample sizes. Lord and Novick (1968), Hambleton (1989), and Thissen and Steinberg (1986) provide technical descriptions of current IRT models. Rules of thumb for the minimum number of examinees required for accurate parameter estimation range from 200 for the one-parameter model (Wright & Stone, 1979), to 500 (Hulin, et al., 1982) or 1,000 (Ree & Jensen, 1980) for the two-parameter model, to 1,000 (Lord, 1968) or 10,000 (Thissen & Wainer, 1982) for the three-parameter model. Hambleton (1989) pointed out that these rules of thumb are not absolute and acknowledged the need for further research in this area.

Barnes and Wise (1991) examined the effectiveness of a modified one-parameter model for estimating ability and item parameters from small samples. Their monte-carlo simulation study showed that incorporating a constant non-zero lower asymptote (guessing parameter) into the one-parameter model resulted in more accurate parameter estimation from small samples than did the traditional one-parameter or three-parameter models. Their results indicated that the modified one-parameter model allowed for accurate estimation of the item difficulty parameters from as little as 50 (simulated) examinees on a 25-item test. However,

their results using the three-parameter model indicated that accurate estimation of the item discrimination and pseudo-guessing parameters could not be obtained from a sample size as large as 200 examinees.

The previous research investigating the effect of sample size on IRT parameter estimation has been performed, for the most part, on simulated data. These studies are significant in that they have demonstrated that recovery of known item parameters from small samples is problematic. However, because real test data were not used, it is not known whether the simulated data accurately reflected the characteristics of small sample data encountered in practice. Furthermore, the results of these studies offer little guidance for test developers who must continue to practice in small-sample settings, regardless of the psychometric accuracy of their item analyses.

A recent study by Stone and Lane (1991) used real test data to investigate the stability of item parameter estimates over time. Using general and restricted forms of the two-parameter model, the authors found that the IRT item parameter estimates remained relatively stable over a one-year time period. Though their investigation did not explore item parameter invariance with respect to sample size (their analyses were based on groups of more than 2,700 examinees), it did demonstrate the utility of using restrictive IRT modeling to evaluate item parameter invariance.

The use of restrictive IRT modeling has also been used successfully to investigate differential item functioning (e.g.,

Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinberg, & Wainer, 1988; in press), and differential testlet functioning Wainer, Sireci, & Thissen, 1991). As with the Stone, et al. (1991) study, the common theme in these studies is the comparison of models where the item parameters are constrained to be equal among groups (restricted models) to models where the item parameters are estimated separately for each group (unrestricted or general models). The two (or more) competing models are compared statistically to determine whether separate estimation of the item parameters for each group (unrestricted model) adds substantially to the fit of the data. If the unrestricted model prevails, the item parameters are not equal for all groups, and item parameter invariance is not exhibited.

Methodology of the Current Study

The present study borrows the IRT model-based procedure used in the Stone et al. (1991), Wainer, et al. (1991), and etc., studies to evaluate the stability of item parameters estimated from a real-life, small-sample testing application. This model-based testing procedure involved obtaining a common set of items administered to three small-sample groups of examinees over a three-year period (sample sizes of 173, 106, and 149, respectively), and estimating the item parameters with both restricted and unrestricted IRT models. The restricted IRT models reflected item parameter stability, where item parameters were constrained to be equal across samples. The unrestricted models reflected item parameter instability and required estimating the

item parameters separately for each group. Thus, the unrestricted models required the estimation of many more parameters than the restricted models.

Evaluating model fit using -2loglikelihoods

All IRT analyses reported here were conducted using the MULTILOG (Version 6.0) IRT software program (Thissen, 1990). MULTILOG is a very general IRT program that fits a variety of IRT models to test data based upon the principal of marginal maximum likelihood (Bock & Aitken, 1981). One index provided by MULTILOG is "negative twice the loglikelihood" (-2loglikelihood) which describes the data's fit to the model (the likelihood of obtaining the data given the model). The -2loglikelihood statistic is the index used to compare competing (i.e., hierarchical) IRT models. Because the difference between the -2loglikelihoods of two hierarchical models is distributed as chi-square, this difference statistic can be evaluated for significance (with degrees of freedom equal to the difference in the number of free parameters estimated in each model). If the additional parameters in the more general (unrestricted) model adds substantially to the data-model fit, then the chi-square difference test will yield a significant value. However, if the chi-square value is not significant, then the more parsimonious (i.e, restricted) model should be accepted. In this study, acceptance of the more restrictive models (insignificant chi-square values) will constitute acceptance of item parameter stability. Acceptance of the unrestricted models will constitute item parameter instability.

IRT Models Tested

Because of the small sample sizes used in this study, it was predicted that only the one and two-parameter models would be appropriate for these analyses. However, for the sake of completeness, the performance of the three-parameter model was also evaluated. The equations for the one, two, and three-parameter logistic models are provided in equations 1, 2, and 3, respectively:

$$P(\theta) = \frac{1}{1+\exp[-\bar{a}(\theta-b)]} \quad (1)$$

$$P(\theta) = \frac{1}{1+\exp[-a(\theta-b]} \quad (2)$$

$$P(\theta) = \frac{c+(1-c)}{1+\exp(-a(\theta-b)]} \quad (3)$$

where P(θ) is the probability of choosing the correct answer as a function of ability θ; *b* is the item's difficulty (expressed in the same metric as ability), *a* is proportional to the slope of the item characteristic curve (ICC) at its steepest point, and *c* is the lower asymptote of the ICC. The item parameters *a*, *b*, and *c* are commonly referred to as the discrimination parameter, difficulty parameter, and guessing parameter, respectively (ä is constant across items and represents the constant discrimination value in the one-parameter model). More detailed descriptions of these, and other, IRT models can be found in Lord and Novick (1968), Hambleton (1989), and Thissen and Steinberg (1986). The restricted IRT models used in this study involve constraining the *a*, *b*, and *c*

parameters for to be equal for identical items taken by the three different groups.

Instrument

The test analyzed in this study is a national certification examination administered to qualified Certified Public Accountants (CPAs) who desire certification in personal financial planning. This examination consists of 100 multiple-choice items and one or more essays. The essay portions of the examination were not included in this analysis[2]. The examination is administered to roughly 150 candidates per year. A subset of 28 items was selected for analysis because these items were reused on three consecutive administrations. Sireci (1991) demonstrated that the ability level of the candidate population over this three-year period was virtually equivalent.

Procedure

The data from three administrations of the examination were combined for the 28-item subset providing a sample size of 428 examinees (106 in 1990, 149 in 1989, and 173 in 1988). This aggregation of data from the three administrations allowed for a larger sample size from which more accurate item parameters could be estimated. The item parameters estimated from the aggregated group were taken to represent the "true" item parameters to which

---

[2]The decision to omit the essay portion of the examination was based on practical reasons, not because of an inability to model the data. An attractive feature of the MULTILOG program is that it allows for mixed-model runs where both dichotomous and categorical items can be analyzed simultaneously. The essays were omitted from analysis because different essays were used in each year of administration.

the sample item parameters would be compared. Descriptive statistics for the 28-item subset for the aggregate group and the three samples are provided in Table 1.

Table 1: Descriptive Statistics for 28-Item Subtest

	Aggregate	1990	1989	1988
N	428	106	149	173
KR20	.47	.51	.49	.41
P	.73	.74	.75	.71
X	20.44	20.72	21.00	19.88
r_b	.37	.39*	.38*	.35

Key: P = Average item difficulty
X = Average score on 28 items
r_b= Average biserial correlation among 28 item subset
* This average based on only 27 items due to a p-value of 1.0 for item 25 in these groups.

Data Analysis: Stage 1

Initially, a classical reliability analysis was performed on the aggregate group (AGG) and the three samples using the SPSS-X procedure RELIABILITY. This analysis was performed to obtain the p-values (proportion of examinees getting each item correct) for each item and a KR20 (internal consistency reliability estimates) value for each group. Because procedure RELIABILITY does not compute biserial correlations, the PRELIS computer program (Joreskog & Sorbom, 1988b) was used to compute the biserial correlation of each item with the total 28-item score.

Following the classical item analyses, traditional one, two,

and three-parameter IRT analyses were performed on the aggregate and sample groups. The -2loglikelihoods of the aggregate models were compared to see which general model best fit the aggregated data. The item parameter estimates obtained from the aggregated group served as the referent parameters to which the parameters obtained from the sample group analyses were compared.

The next stage of analysis involved incorporating the group membership into the IRT modeling. The data was restructured so that each group represented 28 items on a 84-item test (3 X 28=84). Therefore, the restricted modeling involved constraining the item parameters (a, b, and/or c) for items 1 through 28 to be equal to the item parameters for items 29 through 56, and items 57 through 84. The unrestricted modeling involved computing the item parameters by treating the data as comprising a single, 84-item test (each group having missing data on 56 of the 84 items). The unrestricted modeling allowed for separate calibration of the item parameters for each group simultaneously, while the restricted modeling constrained the parameters to be equal for each group.

Results: Stage 1:

Classical item analyses

The p-values and biserials obtained in the classical item analysis indicated that the majority of the items were easy (average p-values ranged from .71 to .75 for each group) and that there was wide variation among the biserial values. The values of these indices are presented in Table 2. The correlations among the p-values and biserials between the aggregate and sample groups are

presented in Table 3. The high correlations among the p-values in Table 3(a) indicate that the p-value index of item difficulty is stable with respect to the three sample groups. Thus, the classical item difficulty parameter was not adversely affected by the relatively small sample sizes. The correlations among the biserial values (Table 3(b)) are much lower than the p-value correlations; especially those correlations among the sample groups. This finding indicates that the biserial index is more sensitive to sample size than is the difficulty index.

Table 2: P-Values and Biserials for 28-Item Subtest (Decimals Omitted)

	AGG		1990		1989		1988	
Item	P	r_b	P	r_b	P	r_b	P	r_b
1	96	41	98	52	97	56	94	24
2	86	33	84	41	89	31	86	26
3	81	42	81	52	79	46	82	35
4	82	31	92	11	87	35	73	28
5	83	28	75	06	91	35	82	36
6	83	23	83	24	85	17	80	25
7	75	45	75	57	71	43	77	43
8	68	24	62	22	67	38	73	18
9	89	43	91	43	91	33	86	47
10	39	39	42	50	46	32	32	32
11	17	33	19	27	13	32	19	41
12	56	29	59	31	52	26	56	32
13	77	34	77	45	75	28	79	35
14	83	41	79	37	86	28	83	55
15	79	20	73	29	83	18	78	13
16	45	28	43	14	45	39	46	28
17	58	41	50	56	58	51	64	28
18	69	61	83	64	67	63	62	58
19	87	39	84	50	88	40	88	30
20	96	66	94	57	97	77	95	57
21	68	44	89	47	87	54	38	33
22	86	35	83	32	87	30	88	43
23	51	40	54	56	56	31	44	33
24	60	34	58	45	66	37	55	20
25	99	36	99	--	99	--	98	31
26	65	36	63	19	69	46	62	36
27	91	39	95	28	89	39	89	43
28	78	44	82	54	77	31	76	51

Table 3: Correlations Among the Classical Item Indices

(a) P-values

	AGG	1990	1989	1988	
AGG	---				
1990	.95**	---			
1989	.97**	.95**	---		
1988	.95**	.82**	.86**		Average $r_{(p)}$ = .92

(b) Biserials

	AGG	1990	1989	1988	
AGG	---				
1990	.73**	---			
1989	.79**	.45*	---		
1988	.75**	.34	.37		Average $r_{(b)}$ = .57

* Significant at p<.01
** Significant at p<.001

IRT Analyses

Assessment of Fit. The one, two and three-parameter models were fit to the aggregated group to determine the most appropriate model for the data. The chi-square difference testing of the -2loglikelihoods indicated that the two-parameter model exhibited the best fit (see Table 4)[3]. This finding is not surprising given

[3]The extremely high probability value obtained for the 2PL versus 3PL test is most likely due to two factors: the 2PL being the best maximum likelihood estimate for the 3PL solution, and the influence of the prior value of .25 set on the lower asymptotes for the 3PL analysis. This finding follows for all subsequent 2PL-3PL comparisons.

the low level of difficulty of the items (eliminating the need for a lower asymptote) and the wide variation in item discrimination (indicating the need for a discrimination parameter) discovered in the classical item analysis.

Table 4: Results of Fit Tests for the 1,2, & 3-P Models

Model	-2loglike.	# Free Parameters	Difference Chi-Square	df	p
3PL	6433	84			
2PL	6447	56	14	28	.983
1PL	6527	29	80	27	<.001

Assessment of item parameter stability. Though the two-parameter model exhibited the best fit, the model-based testing procedure was used to evaluate the stability of the item parameters in all three logistic models. Table 5 presents the results for the one-parameter model. The difference chi-square test was significant, indicating that separate estimation of the item parameters for each group provided better fit than did the model where the (in this case *b*) parameters were constrained to be equal among the three groups. Though the model testing procedure indicated instability among the item parameters in the one-parameter model, the correlations among the *b* parameters for the aggregate group and the three samples were very high. In fact, these correlations were highly similar to the correlations among the p-values in the classical analysis.

The model testing procedure for the two-parameter model also

indicated instability of the item parameters. A significant chi-square value was obtained in comparing the restricted and unrestricted models, indicating that estimation of the item

Table 5: Results from the One-PL Analyses

Model	-2loglike.	# Free Parameters	Difference Chi-Square	df	p
3 Groups	6300	87			
AGG	6511	31	211	56	<.001

$r_{bj's}$:

	AGG	1990	1989	1988
AGG	---			
1990	.94**	---		
1989	.97**	.94**	---	
1988	.97**	.85**	.90**	---

Average r_{bj} = .93

**Significant at p <.001

parameters separately for each group provided a better fit than did the model constraining the item parameters to be equal among the groups. The correlations among the difficulty (b_j) parameters and the discrimination (a_j) parameters also exhibited instability. In fact, the correlations among the difficulty parameters obtained using the two-parameter model were far lower than those obtained using the one-parameter model. This latter finding could be due to a confounding effect of the estimation of the discrimination parameters on the estimation of the difficulty parameters. Table

6 presents the results from the two-parameter analysis.

The model-testing procedure for the three-parameter model also indicated instability among the item parameters. The results of this test are presented in Table 7. Because of the low item parameter correlations in the two-parameter model, and given the poor fit of the restricted three-parameter model, the correlations among the three-parameter estimates were not calculated.

Table 6: Results of the 2-Parameter Analyses

Model	-2loglike.	# Free Parameters	Difference Chi-Square	df	p
3 Groups	6160	170			
AGG	6408	58	248	112	<.001

$r_{bj's}$:

	AGG	1990	1989	1988
AGG	---			
1990	.68**	---		
1989	.58**	.08	---	
1988	.85**	.48	.29	---

Average r_{bj} = .49

$r_{aj's}$:

	AGG	1990	1989	1988
AGG	---			
1990	.50*	---		
1989	.14	.07	---	
1988	.62**	.61**	.10	---

Average r_{aj} = .34

*Significant at p<.01
** Significant at p<.001

Table 7: Results of the 3-Parameter Analysis

Model	-2loglike.	# Free Parameters	Difference Chi-Square	df	p
3 Groups	6113	254			
AGG	6390	86	277	168	<.001

Discussion of Results of Stage 1

The results of the preceding analyses suggest that IRT is not useful to the small-sample test practitioner because the obtained item parameters did not exhibit stability. There are two conclusions that can be reached at this point: (1) the IRT models are inappropriate in this case because of the small numbers of examinees tested in the sample groups, or (2) the IRT models are inappropriate because the 28-item subtest does not meet the assumptions underlying the IRT models[4]. Because the assumptions underlying the IRT model were not evaluated, the power of IRT modeling has not been given a fair trial, and so the second conclusion cannot be ruled out. Therefore, using 20/20 hindsight, we will now move to investigate the second plausible conclusion.

[4]A third plausible explanation for the results is that item-order differences between the three samples affected the parameter estimation (Zwick, 1991). The item ordering was virtually identical between 1989 and 1990, but different for 1988. To investigate this rival hypothesis, "two-group" analyses were run where the 1989 and 1990 data were combined and fit in a two-group parameter estimation model. The parameter estimates in these analyses were also unstable, and so the results are not reported here.

Investigating the Assumption of Unidimensionality

Dorans (1985) and Hambleton (1989), among others, have explained that the fundamental assumption underlying IRT is that the latent variable underlying the test items is unidimensional. These researchers have demonstrated that the other assumptions of IRT (i.e., local independence) will follow consequentially if the unidimensionality assumption is met. In order to evaluate whether the 28-item subtest is unidimensional, a one-factor confirmatory factor analysis model was fit to the data. This analysis was conducted using LISREL-7 (Joreskog & Sorbom, 1988a) to fit the unidimensional model to the matrix of test item intercorrelations. Because all 28 items represented dichotomous variables, tetrachoric correlations were calculated for the matrix of item intercorrelations.

The results of the LISREL analysis indicated that the 28 items were not measuring one, underlying general factor. The one-factor model yielded a very low coefficient of determination (variance accounted for) of .186, a relatively large value for the root mean square residual error (RMSE) of .125, and extremely high standard errors in the lambda-X matrix (ranging from .372 to .411). Again using 20/20 hindsight, this finding is not surprising, because selection of the 28-items was not made with respect to the item content specifications. Rather, the 28 items were chosen because they were the only items common among all three samples of examinees. Thus, the preceding analyses have failed to demonstrate whether IRT can be used in small-sample testing situations *when the*

data fit the model.

To redress this problem, six items were deleted from the 28-item subtest. These six items were deleted based on their content characteristics and poor biserial correlations with the total subtest score. Four of the deleted items (items 4, 5, 6, and 8) belonged to a content area of the test that exhibited poor internal consistency reliability (Sireci, 1991), another item was deleted because of its p-value of 1.0 in two of the sample groups (item 25), and the sixth item (item 15) was deleted because it was the only item representing a particular content area of the test and it had consistently poor discrimination parameter estimates in the two and three-parameter models. Table 8 presents some descriptive statistics for the 22-item subtest.

Table 8: Descriptive Statistics for 22-Item Subtest

	Aggregate	1990	1989	1988
N	428	106	149	173
KR20**	.49 (.55)	.57 (.63)	.49 (.55)	.44 (.50)
P	.71	.72	.72	.68
X	15.40	15.80	15.80	15.00

Key: P = Average item difficulty
X = Average score on 22 items

** KR20 estimates in parentheses are adjusted using the Spearman-Brown formula to correspond to a 28-item test (for comparison with Table 1).

A one-factor confirmatory factor analysis was performed on the "new" 22-item subtest (i.e., on the 22-item matrix of tetrachoric

correlations). The results of this analysis indicated that the 22 items were measuring a general, underlying latent variable. The coefficient of determination for this one-factor model was .572, the RMSE dropped to .105, the standard errors in the lambda-X matrix ranged from .090 to .091, and the LISREL goodness-of-fit indices were .93 (unadjusted) and .92 (adjusted). Though a lower RMSE and higher coefficient of determination would be desirable, these results were accepted as evidence indicating a unidimensional underlying latent variable. Therefore, a second series of IRT analyses (Stage 2) were performed on the data to examine the stability of the IRT item parameters using the 22-item subtest.

Results: Stage 2

The one, two, and three-parameter models were fit to the aggregated data set for the 22 items, and again the two-parameter model exhibited the best fit. The results of this test are presented in Table 9. Subsequently, the restricted and unrestricted analyses were performed for each of the three IRT models. These results were similar to those obtained from analysis of the 28-item subtest. The one-parameter analysis did not exhibit

Table 9: Fit Tests for the 1,2, & 3-P Models -- 22 Item Test

Model	-2loglike.	# Free Parameters	Difference Chi-Square	df	p
3PL	4236	66			
2PL	4235	44	1	22	.999
1PL	4283	23	48	21	<.001

item parameter stability (a chi-square difference statistic of 624 with 44 degrees of freedom was obtained in comparing the restricted and unrestricted models). Similarly, the two-parameter model did not exhibit stability when comparing the restricted and unrestricted models. However, because the two-parameter model exhibited the best fit to the data, and because of the potentially confounding effects of simultaneously restricting the discrimination and difficulty parameters, further analyses were conducted to determine the appropriateness of the two-parameter model for the 22-item data set.

Two other restricted models were used to evaluate the effectiveness of the two-parameter model: one model restricted only the discrimination parameters (a_js) while keeping the difficulty parameters (b_js) unrestricted, the other model restricted the difficulty parameters while keeping the discrimination parameters unrestricted. These two additional models can not assess the ability of the two-parameter model to provide stable difficulty *and* discrimination parameters; however, they can be used to assess the ability of the two-parameter model to provide stable difficulty *or* discrimination parameters, when controlling for the unstable effects of the other parameter.

The results of the two-parameter analysis are presented in Table 10. As mentioned previously, the most restricted model (where the a_js and b_js were constrained to be equal for the three groups) did not exhibit superior fit over the more general (unrestricted) model. Similarly, the model where only the b_js were

restricted did not exhibit improved fit. However, the model constraining only the harder-to-estimate discrimination parameters did nearly exhibit superior fit (p value for chi-square of 64 with 44 df = .023). These results suggest that the estimation of the discrimination parameters did indeed have an adverse effect on estimation of the difficulty parameters. Furthermore, these results suggest that although the two-parameter model was not appropriate for stable estimation of the discrimination parameters, it may be appropriate for stable estimation of the difficulty parameters.

To further evaluate the stability difficulty parameters obtained in the two-parameter, *aj*-restricted model, the correlations among the difficulty parameters for the aggregate and sample groups were obtained (these results are also presented in Table 10). The correlations for the discrimination parameters are not reported, because they are spuriously inflated due to the restrictions placed upon them. An inspection of Table 10 illustrates that the correlations for the item difficulty parameters, though high, are still less than those obtained using classical p-values, or the one-parameter *bjs* during the Stage 1 analyses.

Table 10: Results of 2-Parameter Analyses -- 22-Item Test

Model	-2loglike.	# Free Parameters	Difference Chi-Square	df	p
3 Groups	3993	134			
AGG (both a_js and b_js constrained)	4213	46	220	88	<.001
a_js only	4058	90	64	44	.023
b_js only	4080	90	87	44	<.001

$r_{bj's}$:

	AGG	1990	1989	1988	
AGG	---				
1990	.80**	---			
1989	.81**	.69**	---		
1988	.54**	.38	.24	---	Average r_{bj} = .58

** Significant at p<.001

The results of the restricted and unrestricted three-parameter analysis on the 22-item subset is presented in Table 11. Not surprisingly, these results indicate that item parameter stability was not obtained using the three-parameter model where all three parameters were constrained to be equal among the groups. Similar to the two-parameter analyses, a semi-restricted model, where constraints were imposed on the a_js and c_js only, provided a better fit to the data than the unrestricted model.

Table 11: Results of 3-Parameter Analysis -- 22-Item Test

Model	-2loglike.	# Free Parameters	Difference Chi-Square	df	p
3 Groups	3996	200			
AGG (a_js, b_js, and c_js constrained)	4217	68	221	132	<.001
a_js and c_js only constrained)	4047	112	51	88	.999

Discussion

So where has our rather lengthy investigation of IRT in small-sample testing led us? Can we conclude that IRT is applicable in small-sample testing applications? The results of this study do not provide an unequivocal answer. However, the results do provide hope that IRT can be used with small-samples in some testing situations.

The results reported here are largely consistent with previous research investigating the effectiveness of IRT models on small data sets. The results suggested that stable item difficulty parameters can be obtained from small sample sizes using the one-parameter or modified two-parameter model ***when the data fit the IRT model*** (i.e., are unidimensional). However, there is no incentive for the small-sample test practitioner to use these models for item difficulty estimation, because the easier-to-understand p-values are just as good.

The results are also consistent with previous research

investigating the stability of item discrimination parameters from small data sets. The IRT and classical analyses performed here could not successfully provide stable item discrimination parameters. Thus, for accurate estimation of an item's discrimination larger numbers of candidates are certainly needed. After all, when small numbers of examinees are tested, it only takes a few smart examinees to get an easy item wrong, or a few not-so-smart examinees to get a hard item correct, for the discrimination parameter to become aberrant. Only through the law of large numbers will such aberrancies wash out.

The preceding paragraphs may lead one to conclude that IRT is not applicable for small samples. There are two factors that argue against this point. The first is that item response data can be aggregated from several small-sample test administrations until a larger, more suitable, sample size is obtained. In the analyses reported here, there was a wide amount of variation among the item discrimination parameters estimated in any given year. However, the discrimination parameters obtained from the aggregated sample of 428 examinees should be relatively stable. Thus, aggregation of data from several small-sample test administrations offers promise for those small-sample test administrators who wish to use IRT. Second, through the use of restrictive IRT modeling, the small-sample testing practitioner can alternately free and fix specific item parameters to obtain the most stable parameter estimates given the particular constraints of her/his situation. For example, if there are several items within a test that have been used several

times previously, the aggregated data, for these items only, could be used in an item analysis focusing on the entire test. The item parameters for the newer items could be constrained to fit a more restrictive (e.g., one-parameter) model, while the parameters for the "larger-n" items could remain unrestricted. The restricted models investigated in this study, are perhaps only the tip of the iceberg for application in small-sample test administrations. Thus, restrictive IRT modeling, where restrictions are placed on the parameters for individual items, also offers promise to the small-sample test practitioner.

The present study was limited in two major respects. First, the test data used in this study, though realistic, probably represent the worst test data that could be modeled with IRT. The test used here was a heterogeneous test administered to a highly homogeneous population of examinees. The heterogeneity of the test content (there are 7 content areas represented by the items) directly impeded the unidimensionality of the instrument. The homogeneity of the examinee population negatively affected the calibration of the item parameters. The examinees can be considered highly homogeneous because, first, they are all CPAs, and second, they all practice within a small area of public accounting (personal financial planning). The low variation among the p-values is further evidence of the homogeneity of this group. Therefore, a better investigation of the utility of IRT in small-sample situations might be in a classroom setting, where school teachers can aggregate data over several small classrooms, or over

several years of students (c.f., Nitko, 1983).

The second limitation of this study is that only the effects of sample size on *item* parameter invariance was investigated. This study did not address the estimation of *person* parameters (i.e., ability estimation). Future research should explore the utility of restrictive IRT models for investigating the stability of IRT ability estimation using small samples.

In conclusion, I will return to the original research question "Is IRT useful for the small-sample test practitioner?" The results of this study suggest that IRT does offer some promise for these practitioners, but they certainly have their work cut out for them.

References

Barnes, L.B., & Wise, S.L. (1991). The utility of a modified one-parameter IRT model with small sample sizes. *Applied Measurement in Education, 4,* 143-157.

Bock, R.D., and Aitken (1981). Marginal maximum likelihood estimation of item parameters: application of an algorithm. *Psychometrika, 46,* 443-459.

Dorans, N.J. (1985). Item parameter invariance: the cornerstone of item response theory. *Research in Personnel and Human Resources Management, 3,* 55-78.

Hambleton, R.K. (1989). Principles and Selected Applications of Item Response Theory. In R.L. Linn (Ed.), *Educational measurement,* (3rd edt.), Washington, DC: American Council on Education, pp. 147-200.

Hulin, C.L., Lissak, R.I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: a Monte Carlo study. *Applied Psychological Measurement, 6,* 249-260.

Joreskog, K.G., and Sorbom, D. (1988a). *LISREL 7 User's Reference Guide.* Mooresville, IN: Scientific Software.

Joreskog, K.G., and Sorbom, D. (1988b). *PRELIS: a program for multivariate data screening and data summarization.* Mooresville, IN: Scientific Software.

Lord, F.M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28,* 989-1020.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Ree, M.J., & Jensen, H.E. (1980) Effects of sample size on linear equating of item characteristic curve parameters. In D.J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference.* Minneapolis: University of Minnesota.

Nitko, A.J. (1983). Item analysis appropriate for domain-referenced classroom testing. University of Pittsburgh Project Technical Report No. 1: Pittsburgh, PA.

Sireci, S.G. (1991). *Equating and Scaling of the 1990 APFS Examination.* Technical Report No. 91-1, Examinations Division, American Institute of Certified Public Accountants, New York, NY.

Stone, C.A., & Lane, S. (1991). Use of restricted item response theory models for examining the stability of item parameter estimates over time. *Applied Measurement in Education, 4,* 125-141.

Thissen D. (1990). *MULTILOG (version 6.0) user's guide.* Mooresville, IN: Scientific Software.

Thissen, D. & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 49, 501-519.*

Thissen, D. & Steinberg, L., Gerrard, M. (1986) Beyond group-mean differences: the concept of item bias. *Psychological Bulletin, 99,* 118-128.

Thissen, D. Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.) *Test Validity.* Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D. & Steinberg, L., & Wainer, H. (in press). Detection of differential item functioning using the parameters of item response models. In P.W. Holland & H. Wainer (Eds.), *Differential Item Functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D. & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47,* 397-412.

Wainer, H. & Mislevy, R.J. (1990). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing: a primer.* Hillsdale, NJ: Lawrence Erlbaum.

Wainer, H., Sireci, S.G., & Thissen, D. (1991). Differential testlet functioning: definitions and detection. *Journal of Educational Measurement, 28,* 197-219.

Wainer, H. & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics, 12,* 339-368.

Wright, B.D. & Stone, M.H. (1979). *Best test design.* Chicago: MESA Press.

Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice, 10,* 10-15.